NCBI Molecular Biology Resources

A Field Guide

June 10, 2003

NCBI Resources

- About NCBI
- NCBI Sequence Databases
  - Primary Database - GenBank
  - Derivative Databases - RefSeq
- Entrez Databases and Text Searching
- BLAST Services
- Genomic Resources

The National Institutes of Health

Bethesda, MD
The National Center for Biotechnology Information

- Created as a part of NLM in 1988
  - Establish public databases
  - U.S. National DNA Sequence Database
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don’t add additional information
    - Example: GenBank

- Derivative Databases
  - Human curated
    - Compilation and correction of data
    - Example: SWISS-PROT, NCBI RefSeq mRNA
  - Computationally Derived
    - Example: UniGene
  - Combinations
    - Example: NCBI Genome Assembly
What is GenBank? NCBI's Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
- GenBank data
  - Direct submissions (BankIt, Sequin)
  - Batch submissions via email (EST, GSS, STS)
  - ftp accounts for sequencing centers
- Collaborating databases:
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory Database (EMBL)

The International Sequence Database Collaboration

- NCBI
- EMBL
- DDBJ
- SRS
- NIG
- NIH

GenBank: NCBI's Primary Sequence Database

- Release 135: April 2003
  - 24,027,936 Records
  - 31,099,264,455 Nucleotides
  - 120,000 + Species

- Full release every two months
- Incremental and cumulative updates daily
- Available only through internet


114 Gigabytes
**GenBank Divisions**

<table>
<thead>
<tr>
<th>Bulk Sequence Divisions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PAT</td>
<td>Patent</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tags (240 files)</td>
</tr>
<tr>
<td>STS</td>
<td>Sequence Tagged Sites</td>
</tr>
<tr>
<td>GSS</td>
<td>Genome Survey Sequences (66 files)</td>
</tr>
<tr>
<td>HTG</td>
<td>High Throughput Genome (58 files)</td>
</tr>
<tr>
<td>HTC</td>
<td>High Throughput cDNA</td>
</tr>
<tr>
<td>CON</td>
<td>Contig</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Traditional Divisions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>Bacterial and Archeal</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrate</td>
</tr>
<tr>
<td>MAM</td>
<td>Mammalian (ex. ROD and PRI)</td>
</tr>
<tr>
<td>PHG</td>
<td>Phage</td>
</tr>
<tr>
<td>PLN</td>
<td>Plant and Fungal</td>
</tr>
<tr>
<td>PRI</td>
<td>Primate</td>
</tr>
<tr>
<td>ROD</td>
<td>Rodent</td>
</tr>
<tr>
<td>SYN</td>
<td>Synthetic (vectors, synth. genes)</td>
</tr>
<tr>
<td>VRL</td>
<td>Viral</td>
</tr>
<tr>
<td>VRT</td>
<td>Other Vertebrate</td>
</tr>
</tbody>
</table>

**The Growth of GenBank**

- Release 135: 24.0 million records
- 31.1 billion nucleotides
- Doubling time = 12 months

**Traditional GenBank Divisions**

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized
A Traditional GenBank Record

GenBank Record: Locus

GenBank Record: Identifiers
GenBank Bulk Sequence

Produced in quantity

low accuracy

poorly characterized

EST Division: Expressed Sequence Tags

make cDNA library ➔ 80-100,000 unique cDNA clones in library
What is UniGene?
A gene-oriented view of sequence entries

- MegaBlast based automated sequence clustering
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents


EST hits A. t. serine protease mRNA

Distribution of 119 Blast Hits on the Query Sequence

A. t. mRNA

5' EST hits

3' EST hits

Arabidopsis UniGene Statistics

- 57,746 mRNAs + gene CDSs
- 85,786 EST, 3' reads
- 41,682 EST, 5' reads
- 35,899 EST, other/unknown

221,113 total sequences in clusters

Final Number of Clusters (sets)

26,813 sets total
- 175,000,000 bp
- 1 known gene
- 25,498 expected genes
- 1 EST
- 5% uncharacterized transcripts
- 17,933 sets contain at least one EST
- 16,892 sets contain both genes and ESTs
### Hs UniGene Statistics

<table>
<thead>
<tr>
<th>mRNAs</th>
<th>HTC</th>
<th>EST, 3' reads</th>
<th>EST, 5' reads</th>
<th>EST, other/unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>106,900</td>
<td>7,500</td>
<td>1,451,825</td>
<td>1,886,373</td>
<td>682,624</td>
</tr>
</tbody>
</table>

- **4,135,222** total sequences in clusters

### Final Number of Clusters (sets)

<table>
<thead>
<tr>
<th>sets total</th>
<th>sets contain at least one known mRNA</th>
<th>sets contain at least one HTC</th>
<th>sets contain at least one EST</th>
<th>sets contain both mRNAs and ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>27,887</td>
<td>5,999</td>
<td>109,660</td>
<td>26,547</td>
<td></td>
</tr>
</tbody>
</table>

### UniGene Collections: Animals

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens (human)</td>
<td>4,135,222 111,064</td>
</tr>
<tr>
<td>Mus musculus (mouse)</td>
<td>3,382,434 86,505</td>
</tr>
<tr>
<td>Rattus norvegicus (rat)</td>
<td>422,489 52,264</td>
</tr>
<tr>
<td>Bos taurus (cow)</td>
<td>209,914 18,003</td>
</tr>
<tr>
<td>Sus scrofa (pig)</td>
<td>69,048 14,771</td>
</tr>
<tr>
<td>Gallus gallus (chicken)</td>
<td>125,827 5,042</td>
</tr>
<tr>
<td>Danio rerio (zebrafish)</td>
<td>217,019 15,707</td>
</tr>
<tr>
<td>Oncorhynchus mykiss (rainbow trout)</td>
<td>59,720 11,366</td>
</tr>
<tr>
<td>Oryzon latipes (medaka)</td>
<td>81,296 6,115</td>
</tr>
<tr>
<td>Xenopus laevis (frog)</td>
<td>177,814 19,045</td>
</tr>
<tr>
<td>Silurana tropicalis (frog)</td>
<td>110,976 7,674</td>
</tr>
<tr>
<td>Ciona intestinalis (tunicate)</td>
<td>464,060 13,328</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>276,922 14,701</td>
</tr>
<tr>
<td>Anopheles gambiæ (mosquito)</td>
<td>91,352 3,243</td>
</tr>
<tr>
<td>Caenorhabditis elegans (nematode)</td>
<td>200,436 19,045</td>
</tr>
</tbody>
</table>

### UniGene Collections: Plants and Others

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana (thale cress)</td>
<td>221,113 26,813</td>
</tr>
<tr>
<td>Glycine max (soybean)</td>
<td>201,710 9,084</td>
</tr>
<tr>
<td>Medicago truncatula (barrel medic)</td>
<td>97,063 5,729</td>
</tr>
<tr>
<td>Vitis vinifera (wine grape)</td>
<td>63,652 7,736</td>
</tr>
<tr>
<td>Lycopersicon esculentum (tomato)</td>
<td>70,118 3,749</td>
</tr>
<tr>
<td>Solanum tuberosum (potato)</td>
<td>33,852 2,899</td>
</tr>
<tr>
<td>Tritium aestivum (wheat)</td>
<td>297,961 20,376</td>
</tr>
<tr>
<td>Hordeum vulgare (barley)</td>
<td>263,407 10,272</td>
</tr>
<tr>
<td>Zea mays (maize (corn))</td>
<td>163,541 13,512</td>
</tr>
<tr>
<td>Oryza sativa (rice)</td>
<td>94,780 19,191</td>
</tr>
<tr>
<td>Sorghum bicolor (sorghum)</td>
<td>17,742 802</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii (alga)</td>
<td>80,771 5,712</td>
</tr>
<tr>
<td>Dictyostelium discoideum (slime mold)</td>
<td>119,140 4,782</td>
</tr>
</tbody>
</table>
**Genome Sequencing - HTG, GSS, (WGS)**

Whole BAC insert (or genome)  

- Shredding  
- Sequencing  
- Cloning/Isolating  
- GSS Division or Trace Archive  
- Assembly  

**Draft Sequence (HTG division)**

**GSS Division: Genome Survey Sequences**

- Genomic equivalent of ESTs  
- BAC and other first pass surveys  
- BAC end sequences  
- Whole Genome Shotgun (some)  
- RAPIDS and other anonymous loci

**GenomicClone (BAC)**

gbdiv_gss[Properties]

**Working Draft Sequence**

gaps
HTG Division: High Throughput Genome

Zebrafish PAC Clone

phase 1
Acc = AC109580.1

phase 2
Acc = AC109580.11

phase 3
Acc = AC109580.14

VR1

40,000 to > 350,000 bp

HTG Division: High Throughput Genome

NCBI

Whole Genome Shotgun

40,000 to > 350,000 bp
Whole Genome Shotgun in GenBank

STS Division: Sequence Tagged Sites

- Segment of gene, EST, mRNA or genomic DNA of known position (microsatellite)
- PCR with STS primers gives one product per genome
- Basis of Radiation Hybrid Mapping
  - UniGene
  - Genome Assembly
- Related resource: Electronic PCR
  http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi

RH mapping using STSs

Mouse or Human Chromosome

A  B  C  D

Hybrid Cells

A  B  C  D

PCR Results

A  +  -  +
B  +  -  +
C  -  +  -
D  +  +  -
RefSeq: NCBI’s Derivative Sequence Database

- Curated transcripts and proteins
  - reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis
  - C. elegans
- Model transcripts and proteins
- Assembled Genomic Regions (contigs)
  - draft human genome
  - mouse genome
- Chromosome records
  - microbial
  - organelle

RefSeq Benefits

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators
RefSeq Accession Numbers

<table>
<thead>
<tr>
<th>mRNAs and Proteins</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_123456</td>
<td>Curated mRNA</td>
</tr>
<tr>
<td>NP_123456</td>
<td>Curated Protein</td>
</tr>
<tr>
<td>NR_123456</td>
<td>Curated non-coding RNA</td>
</tr>
<tr>
<td>XM_123456</td>
<td>Predicted mRNA</td>
</tr>
<tr>
<td>XP_123456</td>
<td>Predicted Protein</td>
</tr>
<tr>
<td>XR_123456</td>
<td>Predicted non-coding RNA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Records</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NG_123456</td>
<td>Reference Genomic Sequence</td>
</tr>
<tr>
<td>NT_123456</td>
<td>Contig</td>
</tr>
<tr>
<td>NC_123456</td>
<td>WGS Supercontig</td>
</tr>
<tr>
<td>NC_123455</td>
<td>Chromosome</td>
</tr>
</tbody>
</table>

Useful Entrez Limits: RefSeqs

<table>
<thead>
<tr>
<th>Query Limit</th>
<th>Accession Prefixes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>srcdb_refseq_reviewed[prop]</td>
<td>NC_, NG_, NM_, NP_</td>
<td>Reviewed</td>
</tr>
<tr>
<td>srcdb_refseq_provisional [prop]</td>
<td>NC_, NP_</td>
<td>Provisional</td>
</tr>
<tr>
<td>srcdb_refseq_model[prop]</td>
<td>XM_, XP_, XR_</td>
<td>Model</td>
</tr>
</tbody>
</table>

Reviewed Transcript: srcdb_refseq_reviewed[prop] AND biomol_mrna[prop]

GenBank Sequences: Human Lipoprotein Lipase
Curated RefSeq Records: NM_, NP_

DEFINITION
Mucoraceae lipoprotein lipase (Lpl), human.

NM_008509
XM_134193

AA change
XM_134193

Alignment Based Models

Spidey is an in silico search program that compares the nucleotide and amino acid sequences of human genes in the RefSeq database. The program uses multiple alignment and blast methods to predict the encoded protein sequence from the DNA sequence.

Alignment Based Models

Exon 7: 18915-19035 (genomic); 1218-1338 (mRNA)

XM_134193

NM_008509
Third Party Annotation (TPA) Database

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt

TPAs in Entrez

- Genome annotations
- Genes
- Transcripts
- EST assemblies
Sample TPA records

mouse genomic annotation

corn est assembly

NCBI Web Access

- BLAST
- VAST
- Entrez
- Text
- Sequence
- Structure

Some Web Statistics

- BLAST alone: currently 80,000 searches per day
- Entrez (all databases): 4,000,000 searches per day

Web Access

- Text
- BLAST
- Sequence
- Structure

Some statistics about users per weekday

- 0 5000 10000 15000 20000

- BLAST alone: currently 80,000 searches per day
Users per day

Using Entrez
An integrated database search and retrieval system
NCBI Entrez is a retrieval system for searching several basic databases, it provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide database (GenBank)
- Protein sequence database
- Structure: Three-dimensional macromolecular structures
- GenBank: Comprehensive genome sequence database
- RefSeq: Reference sequences
- SWISS-PROT: Protein sequences
- PIR: Protein Information Resource
- PRF: Protein Research Foundation
- PDB: Protein Data Bank
- OMIM: Online Mendelian Inheritance in Man
- Genomes: Organisms in NCBI sequence databases
- MMDB: Experimental 3D structures
- CDD: Conserved protein domains
- 3D Domains: Compact 3D protein domains in MMDB
- SNP: Single nucleotide polymorphisms
- UniSTS: Sequence Tagged Site markers
- ProbeSet: Gene expression and microarray datasets
- PopSet: Population study datasets
- UniGene: Gene-based expressed sequence clusters

Database Searching with Entrez

- Using limits and field restriction to find plant g6pdh
- Linking and neighboring with g6pdh
### Adding Terms: Preview/Index

**Accession**

**All Fields**

**Preview**

**Index**

Add Term(s) to Query or View Index:
- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.
- Multiple terms selected from Index will be ORed, click AND to add to search.

<table>
<thead>
<tr>
<th>Organism</th>
<th>green plants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filter</strong></td>
<td></td>
</tr>
<tr>
<td>green shirts(1)</td>
<td></td>
</tr>
<tr>
<td>green rice leafhopper(1)</td>
<td></td>
</tr>
<tr>
<td>green rock lobster(2)</td>
<td></td>
</tr>
<tr>
<td>green sea turtle(10)</td>
<td></td>
</tr>
<tr>
<td>green striped grasshopper(5)</td>
<td></td>
</tr>
</tbody>
</table>

Click OR NOT to add terms selected from Index to the query box.

### Plant g6pdh mRNAs

**Summary**

Brief

ASB1

FASTA

ASN.1

XML

GI list

LinkOut

Nucleotide Neighbors

Gene Links

Probe Set Links

Database Links

SNP Links

Protein Links

Structure Links

Taxonomy Links

UniSTS Links

**Links and neighbors**

(related records)

### Alfalfa G6PDH

**Line**

FASTA Format

FASTA Definition Line
>gi|603218|gb|U18238.1|MSU18238

DATABASE IDENTIFIERS

genbank GenBank

FASTA Format

FASTA Definition Line
>gi|603218|gb|U18238.1|MSU18238

DATABASE IDENTIFIERS

genbank GenBank

Abstract Syntax Notation: ASN.1

seq-entry : = seq {
  level 1 ,
  class nucl ,
  desc { title "Medicago sativa glucose-6-phosphate dehydrogenase mRNA, and translated products" ,
        source { 
          taxname "Medicago sativa subsp. sativa" ,
          db { 
            db "taxon" ,
            tag "id 56147" } } }
}

GenPept GenBank

Nucleotide

Protein

GenBank
Protein Neighbors and Links

NCBI Protein Neighbors and Links

Protein Neighbors: Related Proteins

Protein Neighbors->PopSet Links

The population genetics of the origin and divergence of the Drosophila simulans complex species. (Kliman R. M. et al., direct submission)
Genes Database: All Genomes

Entrez Structures

MMDB: Molecular Modeling Data Base

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation, and other information
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)
Structure Summary

Casein kinase S. pombe
Src Kinase H. sapiens
Conserved ATP binding site
Conserved Domains
3D Domain Neighbors
Cn3D viewer
Structure Neighbors
Cn3D 4.1
Cn3D 4.1: Structural Alignment
Cn3D: Simple Homology Modeling

Using Cn3D to model domains

Intermission